

DOCUMENT RESUME

ED 223 641

TM 820 673

AUTHOR Mead, Ronald J.
TITLE Some Useful Ways of Looking at Fit With the Simple Logistic Model.
SPONS AGENCY National Board of Medical Examiners, Philadelphia, Pa.
PUB DATE 20 Mar 82
NOTE 15p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, March 20-22, 1982). Paper also supported by the Intran Corporation, Minneapolis.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Analysis of Variance; *Estimation (Mathematics); *Goodness of Fit; Group Testing; Item Analysis; *Latent Trait Theory; *Mathematical Models; Psyc..ometrics; Scores; *Statistical Analysis; Statistical Bias
IDENTIFIERS *BICAL Computer Program; *Rasch Model

ABSTRACT

The psychometric properties associated with the "specific objectivity" of the simple logistic model are acknowledged to be far superior to those of competing models. It is therefore the ideal model to be used for test development or for the analysis of any measuring instrument developed within a well-defined psychological construct. In estimating the model's parameters, relevant questions addressed are whether the data conform to the model well enough to be useful and how to protect estimates from disturbances to the measurement process specific to individual subjects and items. Fit analysis is considered within the central model concept of Rasch's specific objectivity; parameter estimates are invariant over particular situations. Alternative fit statistics methods are discussed, particularly the chi-square using biased and unbiased condition estimates. The CALFIT program and development of the BICAL program using chi square fit statistics are described. The total unweighted fit statistic in between-score group statistics considers the fundamental data matrix of fit analysis. Between-groups analysis should account for possible subsets that are potentially dangerous to measurement. No one statistic is necessary for every analysis and none by itself is sufficient to ensure specific objectivity which depends on analyses sensitive to possible disturbances. (CM)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED223641

***** DRAFT *****

SOME USEFUL WAYS OF LOOKING AT FIT
WITH THE SIMPLE LOGISTIC MODEL

Ronald J. Mead

Educational Measurement Systems, Inc.
112 East Elmwood Place
Minneapolis, MN 55419
(612) 871-6292

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

✓ This document has been reproduced as
received from the person or organization
originating it.
☐ Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official NIE
position or policy.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

R.J. Mead

March 20, 1982

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

This paper was prepared for
The National Council on Measurement in Education
1982 Annual Meeting, New York City

With the support and assistance of
The National Board of Medical Examiners, Philadelphia, PA,
and Intran Corporation, Minneapolis, Minnesota

Tm 820 673

SOME USEFUL WAYS OF LOOKING AT FIT WITH THE SIMPLE LOGISTIC MODEL

It is generally acknowledged that the psychometric properties associated with the "specific objectivity" of the simple logistic model are far superior to those of other competing models. It is therefore the ideal model to be used for test development or for the analysis of any measuring instrument that was carefully developed in accordance with a well defined psychological construct. Because of these desirable properties, it is worth considerable effort to structure the measurement situation in order to employ this model even if data do not fit it exactly. This will without a doubt lead to better instruments, better understanding of the constructs involved, and better educational measurements. It probably also means more work in test development.

Questions associated with estimating the model's parameters (when the data fit) have been solved for all practical purposes (Wright and Panchapakesan, 1969; Andersen, 1972; Gustaffson, 1979) and techniques for applying the results (when the data fit) have been presented (Wright, 1977; Wright and Stone, 1980). The remaining question of how to tell if data do fit is also easily answered: real data never fit the simple ideal model. This is as true of physics as it is of education.

Since the question of fit always has the same answer, it is not of much use as stated. The more relevant questions in practice are

- (1) Do the data conform to the model well enough to be useful?
- (2) How do we protect ourselves from disturbances to the measurement process specific to individual persons and items?

The idea underlying the fit analysis is the central concept of the model itself: Rasch's specific objectivity. All the possible statistics ask in one way or another if the parameter estimates are invariant over a range of conditions that is useful to us. If they are, we have achieved our goal; if not, we have the opportunity to learn something about the process that we did not know when we started.

Alternative Fit Statistics

A variety of fit statistics have been advanced at one time or another for one purpose or another. These include likelihood ratio tests, Pearson chi-squares, discrimination indices, between group mean squares, Z's and T's and total mean squares, Z's and T's both weighted and unweighted, item fit statistics and person fit statistics. These are all useful for some pur-

poses but not for others. Some are alternative approaches or representations of the same thing. It is essential to recognise which is relevant to a particular situation and how it should be judged. The decisive issue is not the development of the definitive statistic (no one statistic is ever necessary or sufficient). The important point is understanding the nature and intent of the process.

Before getting involved in distinctions among statistics, there are two issues involved with the use of the chi square and its derivative statistics that should be mentioned; one well known and perhaps over-emphasized and the other virtually unknown but perhaps more important.

The chi square is computed as the square of the sum of x 's minus the sum of P 's, summed over some relevant groups of people and items, divided by the same summation of P times one minus P . If all the P 's were equal, This would be readily recognisable as a test about the estimate of P for a binomial model.

The first problem is that this calculation assumes that the P 's are known. In that case, $(x-P)/\sqrt{P(1-P)}$ would be a "z--statistic". In practice, the P 's are computed from the estimates of ability and difficulty and, for most forms of the fit statistics, it is not clear what adjustment should be made to the

degrees of freedom.

Secondly, the estimates of ability and difficulty used in computing the P 's must be those which make the sum of x 's exactly equal to the sum of P 's when the data fit the model exactly. This is true when the summation is over the calibrating sample and the estimates are those which satisfy the unconditional maximum likelihood estimation equations. However, these estimates are biased and the final step in every unconditional estimation procedure is to remove this bias. Now we have the right answer but the likelihood equations are no longer maximized and the sum of x 's no longer equals the sum of P 's. Using unbiased estimates introduces a spurious component into the chi square. If these statistics are used, the fit computations should be done with biased estimates regardless of how they were obtained in the first place.

A Brief History of Fit in BICAL

Since the BICAL program evolved concurrently with my thinking about fit, it has passed through several stages which has resulted in some confusion and some inappropriate recommendations (George, 1979; Douglas, 1981) about what it does and does not do. I am concentrating on the chi square statistics that BICAL uses because it is widely used in practice but I do

not intend to minimize the theoretical importance of the work done on likelihood ratio tests by Gustafsson, Fischer and Andersen among others. By theoretical importance I mean the more we know about it the better off we will be in general but I don't expect these statistics to be the answer to the concerns of the practitioner.

BICAL is a direct descendent of MESAMAX and CALFIT. MESAMAX was developed at Chicago by Wright in the mid-60's with help from Panchapakesan and Choppin. Its fit analysis is a between score group mean square based on all score groups from 1 to $L-1$, where L is the number of items. It uses a sum of z -squares, where each z is computed as the standardized difference between the actual number obtaining the correct answer at each score and the number expected given the model and parameter estimates. When summed over all scores and items for reasonable sized samples, this gives a rather good chi square with $(L-1)(L-2)$ degrees of freedom.

CALFIT was produced in the early 70's by Andrich and Penny Edaert under Wright's supervision. It was intended to run on a small computer and to handle small data sets. In order to save space and to avoid cells with low expected counts, it combines adjacent scores to obtain groups with reasonable numbers. The number of groups was restricted to six strictly for programming

considerations. The result is a chi square with $(L-1)(6-1)$ degrees of freedom.

These overall fit statistics which address the issue of fit of all the data simultaneously were recognised from the start as not particularly interesting. They were immediately partitioned into components associated with each item. While there is no problem doing the summation separately for each item, the degrees of freedom can not be divided up so easily. Rather than having five degrees of freedom as we would like, the items have an average of $5(L-1)/L$. If there are techniques for dealing with fractional degrees of freedom, the familiar, "analysis-of-variance" approach to counting degrees of freedom is probably inappropriate in this situation where every observation contains a different amount of information.

No sooner was CALFIT working, then I decided to "improve" it. These changes eventually led to a completely new program now known as BICAL. The original version retained the between score group chi square based on six score groups as its principal fit statistic. At this point the only addition to the fit analysis was a "discrimination index" computed by regressing the score group residuals on the average group ability, an approach similar to that suggested by Andersen (1972). This statistic can be interpreted as a single degree of freedom effect for trend

partitioned out of the between group mean square. It also does a reasonable job of recovering "discrimination" in simulated data.

At the next stage, I became dissatisfied with the lack of invariance in the between score group statistic. It often changed substantially when the group boundaries were changed. This led to the ultimate between group analysis with each person being his own group. (In this analysis, none of the groups are unoccupied but they are all very small.) This is now referred to as the "total unweighted fit statistic" in the BICAL output.

While I now have reservations about the utility of "total" fit statistics, it led me through an important step in my own thinking. They caused me to rediscover the fundamental element of fit analysis: the data matrix. Having come face to face with each person-item response as well as its probability and expected value under the model, it was apparent that these elements could be combined in many more ways than just summing over people for an item or grouping people by score. People could be grouped by any interesting criteria to test a variety of hypotheses about the item and items could be clustered by any criteria to test hypotheses about the person. None of these ideas have been adequately incorporated into BICAL.

Variations And Refinements in BICAL'S Statistics

The "total unweighted fit statistic" is a little troublesome because it is very sensitive to extreme responses. A single high ability examinee missing one very easy item is frequently enough to make the mean square large. While this is an unlikely event and occurs fairly rarely, the magnitude of the mean square produced is often not reflective of the quality of the item and out of proportion to the seriousness of the problem. It may be a very interesting educational event for understanding the person but it has little impact on the quality of the calibration.

This situation led Wright to a "weighted" mean square (Mead, 1979) which was computed by weighting each z-square by the amount of information in the observation. This statistic has a much improved distribution but is not powerful against disturbances in the tails (e.g., miskeying and random guessing).

The standard error associated with a mean square is primarily a function of sample size and so determining what constitutes a large value changes from calibration to calibration and what is big for the total is not big for the between. We soon found ourselves, when trying to interpret analyses, subtracting one and dividing by an approximate standard error (given by the square root of 2 over N) to get an approximate z.

The next logical step was to have the computer do the arithmetic and to do it better in the process.

BICAL now applies a cube root transformation,

$$z = ((U^{1/3} - 1) / (s/3) + (s/3),$$

where s is the standard error of U ,

to transform mean squares into pseudo z statistics, symmetric around zero and with variances near one. Based on simulations done for a specific application using large samples and long tests, this results in a very good approximation to the standard normal distribution when the mean square was computed using biased parameter estimates.

Interpreting BICAL Output

The most important step in interpreting BICAL output is to determine if the version you are using reports mean squares or transformed mean squares. If some of the numbers are negative, then the mean squares have been transformed and their mean should be near zero and variance near one. The untransformed mean squares should have a mean of one and a variance about $(2/N)$.

Practitioners invariably want me to tell them what a big number is. Does the fit statistic need to be two, three four, or some other value, standard errors away from the mean before it is serious. I resist answering this question but in data simulated to fit the model exactly, values of the pseudo z as big as four almost never occur. In most applications, they occur frequently. Using four as the criterion almost guarantees that every instance that is identified represents some departure from the model but some will be missed. If two or one and a half is used, some cases will be identified that represent only random variation but very few real disturbances will be overlooked. Which value is used depends on the costs and penalties of the two types of errors.

Selecting some value such as 2 or 4 or 1.96 as a screen for a computer to sort through large data sets is a perfectly reasonable strategy but religiously adopting some magic number as an absolute standard such that any value greater than the standard is considered misfitting and any value smaller is considered fitting is not satisfactory. If the data do not fit the model, simply cutting off the tails of the distribution will not accomplish fit. The result is instead a truncated distribution with misfitting data. Upon replication, there should be less misfit but some fit statistics will almost certainly exceed our absolute standard.

I prefer to examine the empirical distribution first and discard any obvious outliers, without too much concern for the model variance. If there are outliers, they are usually easily diagnosed as miskeyed, having more than one right answer, having no right answer or something equally obvious. These can be eliminated or repaired.

If the variance of the distribution is still too large, the calibrations are "unbiased" in the sense that they are ordered according to the number of people in this sample who answered them correctly but the standard errors provided by the model are underestimates and the benefits of specific objectivity cannot be automatically assumed. Decisions about examinees based on total score should be made cautiously whether they are transformed to logits or not.

This situation usually reflects a lack of homogeneity in the item content that may be unavoidable but calibrating and reporting by subset should help. It may also be useful to rethink the construct being measured and review the item generation procedures.

I feel very strongly that it would be a mistake at this point to abandon the simple logistic model in favor of an item

response model with more parameters. Since it has been my experience that most departures are indications of multidimensionality, I am unwilling to run the risk of absorbing the multidimensionality into a more complex but nonetheless unidimensional model. I am also unwilling to forego the opportunity to understand the nature of the dimensionality that is handed to me by a selective model that objects to departures from its basic propositions.

Many tests currently in use do not meet Rasch's standards of measurement but there are also many existing instruments that do, without qualification, fit the simple logistic model. There are probably more examples in psychology than in education but they do exist in both. All of them, without question, were laboriously developed from carefully conceived constructs by skilled professionals and all, without question, are good instruments. This seems reason enough to use fit to a Rasch model as a goal if not the standard in all test development.

Fit and Specific Objectivity

The purpose of fit analysis is to achieve specific objectivity: "objectivity" meaning that the parameter estimates are invariant and "specific" meaning in particular situations. It is usually stated that the item difficulty estimates are independent

of the person ability parameters. But it means as well that they are independent of all attributes of the persons: race, grade, sex, age, height not just ability. The fit analysis should take account of any possible subsets in the data that are potentially dangerous to measurement. No one fit statistic can be sensitive to all possible disturbances, so it is prudent to construct "between" group analysis for any groups that are of concern and to include some form of the total fit statistic as general, if not particularly powerful, insurance against unanticipated problems.

Achieving objectivity depends not on understanding exactly the null distributions of the statistics but on carefully constructed instruments and carefully conceived analyses that are sensitive to the particular pitfalls in the situation. No one test statistic is necessary for every analysis and none by itself is sufficient to ensure specific objectivity. It does require practical screening procedures and knowledgeable professionals to apply them. Both are readily available.